Introduction to Big Data Assignment-1

Name: Faizan Mulla

Roll No: 21F1003885

Problem Statement

Spin Up a VM and write a python program to count lines of a file placed in GCS.

a. Submit the python file with your code

b. Also, provide the text file containing your output.

Implementation Steps

Step 1: GCP Environment Setup

1. **Create GCP Project**
    - Accessed Google Cloud Console
    - Created new project for the assignment
    - Enabled necessary APIs
        - Compute Engine API
        - Cloud Storage API

2. **Virtual Machine Configuration**
    - Navigated to Compute Engine → VM instances
    - Created new instance with specifications:
        - Name: **ibd-ga1-vm**
        - Region: **asia-south-1**
        - Default machine configuration
        - Standard boot disk

3. **Cloud Storage Setup**
    - Created bucket "**ibd-ga1-bucket**"
    - **Region**: asia-south1
    - Configured with standard storage class

Step 2: Environment Preparation / VM setup

- Click the "SSH" button next to your VM. This opens a browser-based terminal
- In the top right corner, click on "Upload File" and choose the Python script from your computer. The file will be in your home directory now.

Google Cloud SDK Configuration and authenticate:

- Now, in the SSH terminal, run these commands:

```
curl -O https://dl.google.com/dl/cloudsdk/channels/rapid/downloads/google-cloudsdk-xxx-linux-x86_64.tar.gz

tar -xf google-cloud-sdk-xxx-linux-x86_64.tar.gz

./google-cloud-sdk/install.sh
./google-cloud-sdk/bin/gcloud init
./google-cloud-sdk/bin/gcloud auth application-default login
```

- Type "Y" to log in
- Click the link it shows
- Log in with your Google account
- Copy the verification code shown
- Paste it back in the terminal
- Select your project number when asked
- Choose your default region. For me it is: "**asia-south-1-a**"

Create Virtual Environment

- First install required packages:

  ```
  sudo apt-get update
  sudo apt-get install python3-venv python3-pip
  ```

- Create a directory for your project:

  ```
  mkdir ga1
  cd ga1
  ```

- Create a virtual environment

  ```
  python3 -m venv venv
  ```

- Activate the virtual environment:

  ```
  source venv/bin/activate
  ```

- Now install the Google Cloud Storage package:

  ```
  pip3 install google-cloud-storage
  ```

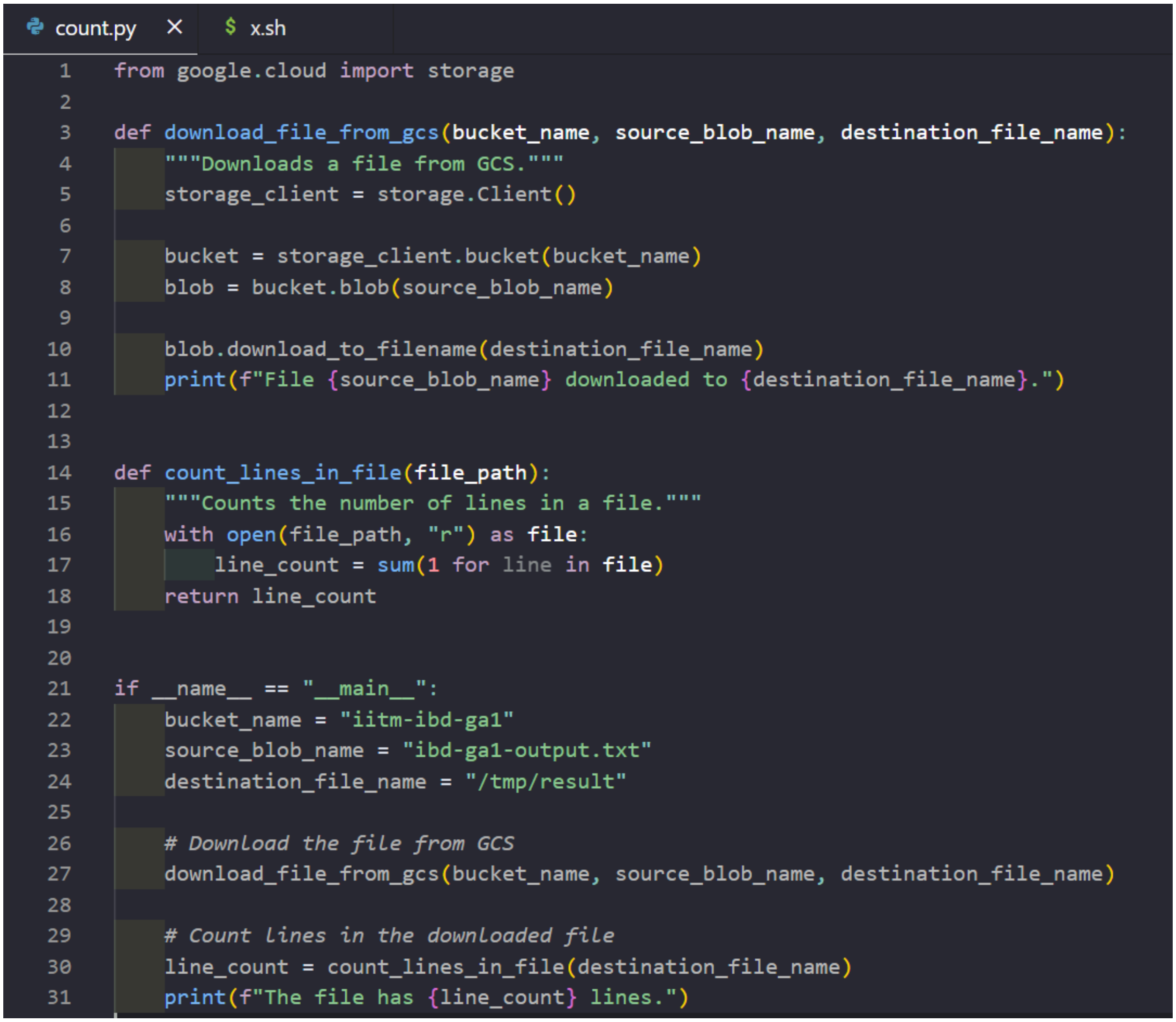
- Check if the Python file is in the project directory or not. If not:

  ```
  cp ~/count.py .
  ```

- **NOTE**: Every time you log into your VM and want to run the script, you'll need to:

  ```
  cd ga1
  source venv/bin/activate
  ```

Step 4: Code Implementation

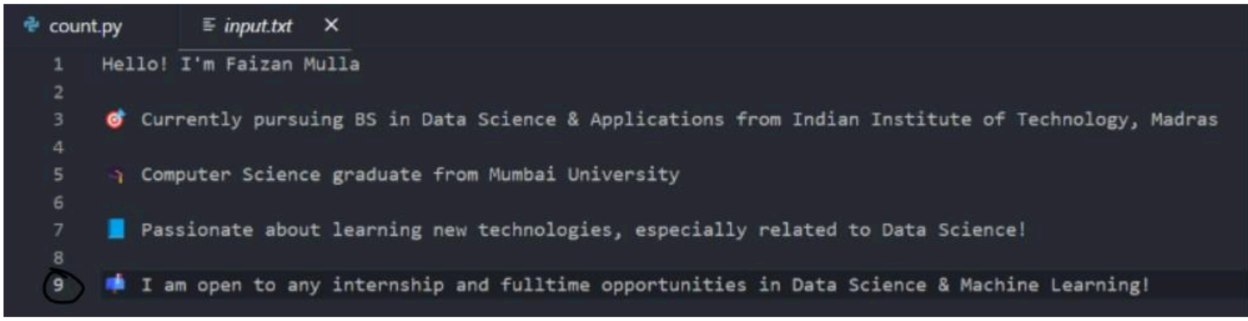


```python
from google.cloud import storage

def download_file_from_gcs(bucket_name, source_blob_name, destination_file_name):
    """Downloads a file from GCS."""
    storage_client = storage.Client()

    bucket = storage_client.bucket(bucket_name)
    blob = bucket.blob(source_blob_name)

    blob.download_to_filename(destination_file_name)
    print(f"File {source_blob_name} downloaded to {destination_file_name}.")


def count_lines_in_file(file_path):
    """Counts the number of lines in a file."""
    with open(file_path, "r") as file:
        line_count = sum(1 for line in file)
    return line_count


if __name__ == "__main__":
    bucket_name = "iitm-ibd-ga1"
    source_blob_name = "ibd-ga1-output.txt"
    destination_file_name = "/tmp/result"

    # Download the file from GCS
    download_file_from_gcs(bucket_name, source_blob_name, destination_file_name)

    # Count lines in the downloaded file
    line_count = count_lines_in_file(destination_file_name)
    print(f"The file has {line_count} lines.")
```

Step 5: Execution and Results

1. Script Execution
    - Ran the Python script on VM using python3 count.py
    - Successfully accessed GCS bucket
    - Processed all files in bucket

2. Results

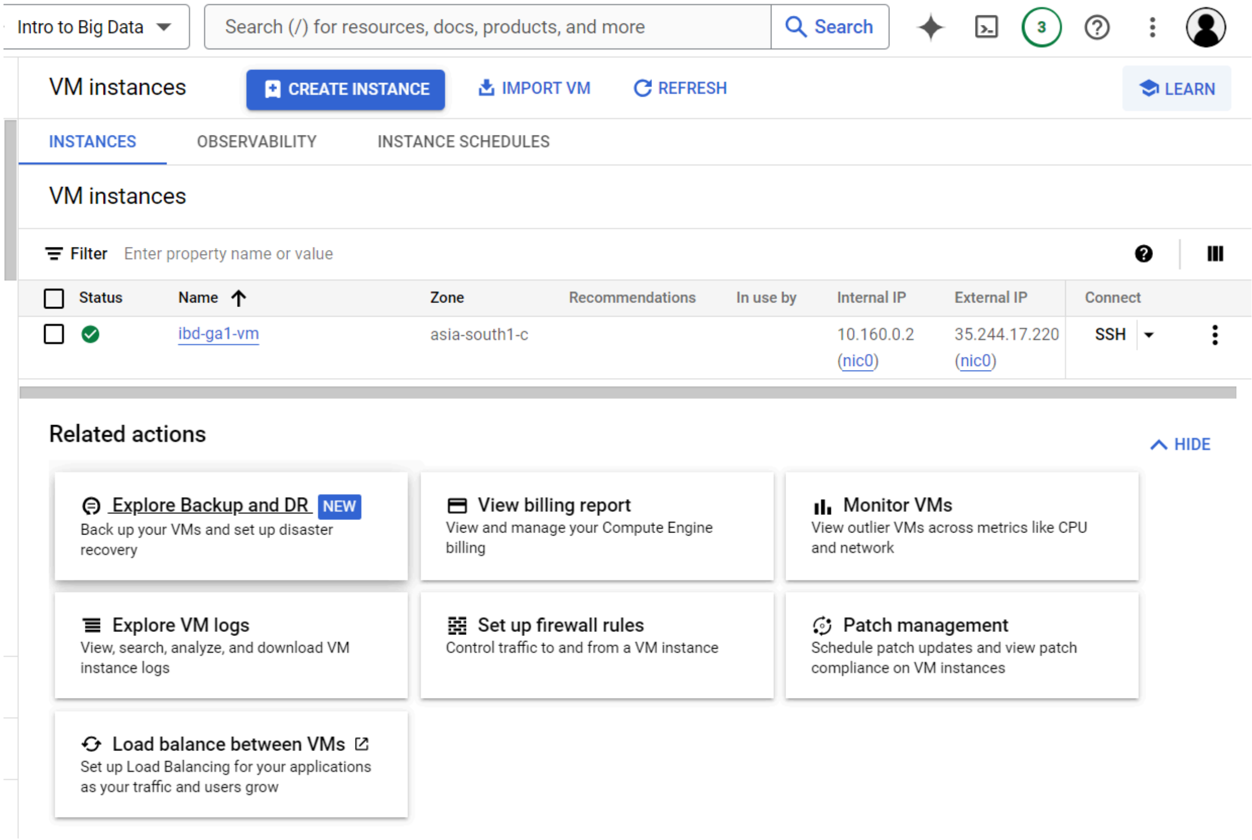


```
1  Hello! I'm Faizan Mulla
2
3  🎯 Currently pursuing BS in Data Science & Applications from Indian Institute of Technology, Madras
4
5  🎓 Computer Science graduate from Mumbai University
6
7  📘 Passionate about learning new technologies, especially related to Data Science!
8
9  📫 I am open to any internship and fulltime opportunities in Data Science & Machine Learning!
```

The file has 9 lines.

Relevant Screenshots

1. Virtual Machine Setup



□	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
□	✓	ibd-ga1-vm	asia-south1-c			10.160.0.2 (nic0)	35.244.17.220 (nic0)	SSH

Related actions

Explore Backup and DR NEW
Back up your VMs and set up disaster recovery

View billing report
View and manage your Compute Engine billing

Monitor VMs
View outlier VMs across metrics like CPU and network

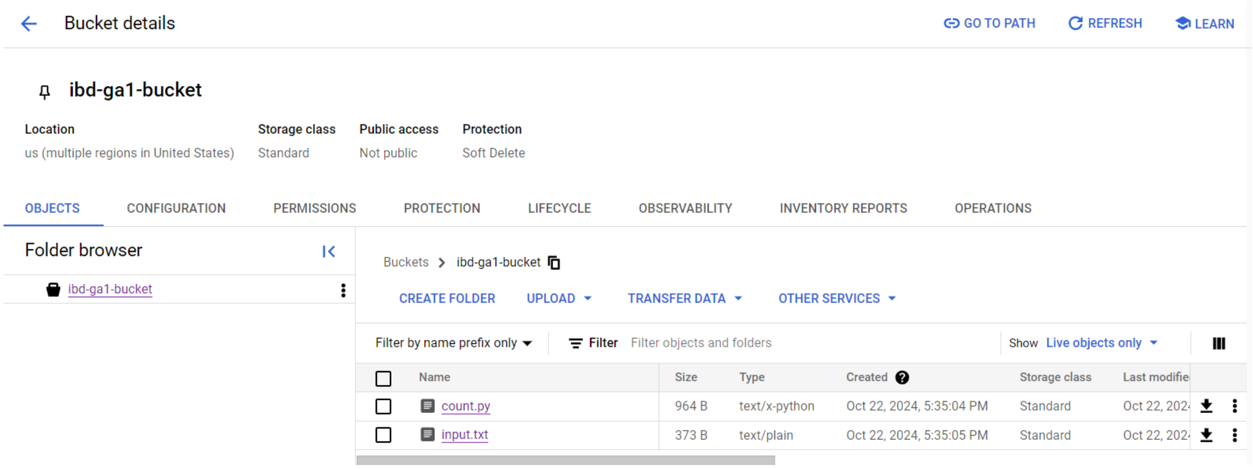
Explore VM logs
View, search, analyze, and download VM instance logs

Set up firewall rules
Control traffic to and from a VM instance

Patch management
Schedule patch updates and view patch compliance on VM instances

Load balance between VMs
Set up Load Balancing for your applications as your traffic and users grow

2. Cloud Storage Bucket

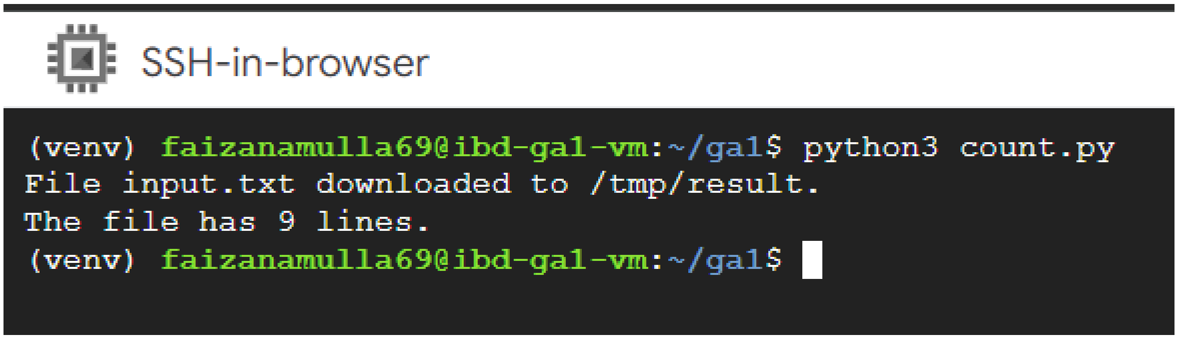


ibd-ga1-bucket

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Soft Delete

	Name	Size	Type	Created	Storage class	Last modifie
□	count.py	964 B	text/x-python	Oct 22, 2024, 5:35:04 PM	Standard	Oct 22, 202
□	input.txt	373 B	text/plain	Oct 22, 2024, 5:35:05 PM	Standard	Oct 22, 202

3. Script Execution



```
(venv) faizanamulla69@ibd-ga1-vm:~/ga1$ python3 count.py
File input.txt downloaded to /tmp/result.
The file has 9 lines.
(venv) faizanamulla69@ibd-ga1-vm:~/ga1$
```